The Trail Making Test: A study of its ability to predict falls in the acute neurological in-patient population

Dr. B.A. Mateen[1,3,6]
MBBS

Mr. M. Bussas[2]
BSc

Dr. C. Doogan[3]
DClinPsy

Dr. D. Waller[4]
MBBS

Dr. A. Saverino[5]
MD

Dr. F. J. Király[2,6*]
PhD

Prof. E. D. Playford[3,7]
MD

[1] Medical School, University College London, London, UK
[2] Department of Statistical Science, University College London, London, UK
[3] Therapy and Rehabilitation Services, National Hospital for Neurology & Neurosurgery, London
[4] Neurorehabilitation Unit, National Hospital for Neurology and Neurosurgery, London, UK
[5] Wolfson Neuro Rehabilitation Centre, St Georges Hospital, London, UK
[6] The Alan Turing Institute, London, UK
[7] Institute of Neurology, University College London, London, UK

Author for correspondence (*):
Dr. Franz J. Király
Department of Statistical Science, University College London,
Gower Street London WC1E 6BT United Kingdom
Tel.: +44 - 20 - 7679 1259 Fax.: +44 - 20 - 3108 3105
E-mail: f.kiraly@ucl.ac.uk

Word Count: 3,314

Key Terms: Accidental Falls, Decision Support Techniques, Trail Making Test, Executive Function, Attention

Abstract

Objective – To determine whether tests of cognitive function, and patient reported outcome measures of motor function can be used to create a machine learning-based predictive tool for falls.

Design – Prospective cohort study.

Setting – Tertiary neurological and neurosurgical centre.

Subjects – 337 in-patients receiving neurosurgical, neurological, or neurorehabilitation-based care.

Main Measures – Binary (Y/N) for falling during the in-patient episode, the Trail Making test (a measure of attention and executive function), and the Walk-12 (a patient reported measure of physical function).

Results - The principal outcome was a fall during the in-patient stay (n = 54). The Trail test was identified as the best predictor of falls. Moreover, addition of other variables, did not improve the prediction (Wilcoxon signed-rank $p < 0.001$). Classical linear statistical modelling methods were then compared with more recent machine learning based strategies, e.g. Random forests, neural networks, support vector machines. The random forest was the best modelling strategy when utilizing just the Trail Making Test data (Wilcoxon signed-rank $p < 0.001$). with 68% (± 7.7) sensitivity, and 90% (± 2.3) specificity.

Conclusion – This study identifies a simple yet powerful machine learning (Random Forest) based predictive model for an in-patient neurological population, utilizing a single neuropsychological test of cognitive function, the Trail Making test.

Introduction

Falls are a serious public health concern with potentially fatal consequences and significant financial implications for individuals,[1-2] their families,[3-4] and the National Health Service.[5] In the UK, falls account for over 60% of all hospital in-patient related safety incidents,[6] with the highest risk of falls most commonly associated with neurological conditions such as the stroke population, in whom the prevalence of falls can be as a high as 50%.[5,7] The issues of falls-related injury is not restricted to the UK; there were more than 10,000 fatal falls in the elderly population, and an additional 2.6 million medically treated falls-related injuries that were non-fatal in the USA in one year alone, resulting in a direct cost of close to US $20 billion.[8] Predicting which patients are at high risk of falling is one of the first steps towards implementing a system to prevent those falls.

The current state-of-the-art methods for predicting falls are based on assessing patient factors such as age, urinary urgency, or walking impairment.[9] By identifying additional risk factors it may be possible to improve the prediction.[10] Several contemporary theories of locomotion have hypothesized the importance of cognitive dysfunction as a risk factor for falling.[11] Specifically, executive function and attention have been shown to be independent falls risk factors.[12-16]

Whilst there are many neuropsychological tests available whose ability to measure executive function is well described in the literature, their relationship to attention is less well understood.[17] In this study we chose to use the Trail Making test because its two parts (A & B) mapped well onto several well accepted theories of attention,[18,19] executive function,[12] and more generally the cognitive control of tasks.[20] The Trail Making test in combination with other variables has been used to predict risk associated with other tasks that rely heavily on executive function and attention, such as driving.[21,22] However, we are

unaware of any studies that have attempted to use the Trail Making Test to create a model for falls prediction in a neurological cohort. The purpose of this study was to determine whether the Trail Making Test, in combination with other risk factor data, is capable of accurately predicting falls in the acute neurological in-patient population.

Methods

Following discussions with our local ethics review committee it was agreed that advice should be sought from the UK statutory regulator, the Health Research Authority, as to whether a full ethics review was needed for the study. The Health Research Authority determined that the appropriate designation for this study was 'Service Development', thus waiving the requirement for ethics board review. The rationale provided was that because the relationship between executive function and falls is well recognized, and thus the study was an application of knowledge, not investigation into an unknown quantity. The study was subsequently vetted and approved by hospital governance. Patient consent was recorded in the clinical notes. Data analysis was conducted on a completely anonymised dataset. Non-anonymized data was stored securely for use by the patient's clinical team, accessible only through the hospitals secured severs. Because the data was used under the service development designation, we are not able to share the raw data. However, if you wish to utilize the model for research purposes, please contact the corresponding author.

Data was collected between the 17th November 2014 and 17th December 2014 at the National Hospital for Neurology & Neurosurgery, Queen Square, UK, a tertiary neuroscience centre, from 3 neurosurgical, 3 neurological, and 2 neuro-rehabilitation wards.

All patients present on the wards at the beginning of the study, and those admitted over the course of the recruitment period, were informed of the project and verbally consented by a trained researcher (Author BAM). Everyone, including those unable to complete the test battery, was included in the study except those patients who did not consent. Contraindications to test administration included: lack of fluency in English; severe cognitive impairment, communication difficulties, or severe mood/behavioral problems where sufficient support could not be clinically provided to allow for fair administration of

the tests; and/or, agreeing to their demographic data being utilized, but declining to complete the test battery.

The test battery (see appendix table S1 for detailed description of tests used) consisted of the Trail Making tests[23,24], a PROM (Patient Reported Outcome Measure) of motor function (Walk-12),[25] and three binary (yes/no) questions relating to the past 1 month's medical history (whether the patient had: undergone surgery; experienced a change in physical function; and/or, fallen), and demographic information (diagnosis, age, sex, ethnicity, & years of education) collected at admission. Testing was carried out by a trained researcher (Author – BAM), under the supervision of a consultant neuropsychologist (Author – DW).

The principal outcome in the prospective study was whether a patient fell or not during their in-patient stay. A fall was defined as an incident, which consisted of unintentional contact with the ground (or intermediary object, which halted their progression to the floor, e.g. a wall), by any part of the body, except the feet. The additional distinction of recurrent falling has been disregarded in this study as a single fall is sufficient to cause injury. Falls are considered serious incidents, and are recorded on a computer-based registry. We used this registry to identify retrospectively which patients recruited into the study fell during their in-patient stay, and matched this information to the prospectively collected data generated by the test battery.

The summary statistics for each test in the battery, in the form of 6 number summaries (minimum, 1st and 3rd quartile, median, mean, and maximum values), are available in the appendix. The mean score for the faller and non-faller populations were then compared using two-tailed t-tests, and the corresponding p-value for significance is reported in the tables.

Next, a series of predictive models were generated to determine which combination of data and statistical model most reliably predicts whether a patient is likely to fall. The modelling was performed using the *R (v 3.2.0)* statistical software suite and the *mlr (v 2.7)* machine learning library.[26,27] Table S2 contains an overview over the different statistical models used. The models considered may be roughly divided into "classical" models such as logistic regression, and "machine learning" methods such as random forests.[28] A random forest can be thought of as a group of slightly different classification trees that are learnt based on the data provided. When new data is then presented to this group of classification trees, each tree uses the new information to arrive at a prediction, in this case, fall or not fall. Each tree's prediction is considered a vote, and the result that the random forest algorithm presents to the user is the class (fall or not fall) that the majority of the trees selected.

The quantitative measures of how reliable each prediction strategy is in predicting new data, is described using the mean misclassification error, sensitivity (= True Positive Rate), specificity (= True Negative Rate), precision (= Positive Predictive Value), and the F1 score (a classical measure of the trade-off between sensitivity and specificity). Using measures of accuracy in isolation, such as sensitivity, can be misleading if a model 'cheats' (i.e. does not use the data to predict outcome, but instead in classification tasks such as this repeatedly guesses the majority class (no fall) to maximize its score on one measure of accuracy – also known as an idiot or uninformed classifier). To prevent a model being selected that does this, the F1 statistics is utilized. A non-zero f1 statistic would suggest that the classifier is not attempting to cheat, and the higher the f1 statistic, the better the classifier.

For each prediction strategy, the quantitative measures of predictive strength were estimated by repeatedly splitting the data into a training sample on which the model is fitted and a test sample, which mimics "new" data, on which the model is tested by comparing the predictions to the true labels (faller vs non-faller), this is called cross-validation. We utilized a 10-fold cross-validation procedure. Therefore, the data was split into 10 parts, with 9 parts being used to train the model, and the 10^{th} portion being treated as new data. An algorithm, known as the Jackknife estimator of variance, was applied to the results from that 10^{th} portion to produce error estimates for each quantitative measure of predictive strength. The process of training on 9 sets, and testing on the 10^{th} was repeated so that all 10 parts play the role of new data once. The 10 different results are then combined to produce a single estimate for the overall predicted model performance, and the associated error statistic. The performance of a strategy was considered better than another if the difference was significant at 5% significance level of a Wilcoxon signed-rank test.

Three Receiver Operator Characteristics (ROC) curves were generated to illustrate the benefit of using the most informative sub-set of data, and secondly, the best modelling strategy for this scenario, compared to the other methods and data available.

Finally, the above analysis was carried out on restricted datasets, utilizing only the subset of individuals without any missing values to allow comparison between subsets of the data, and modelling strategies. Given that the ability to complete all tests is not reflective of the total sample, it needs to be determined whether the chosen predictive approach generalizes to the whole population. To illustrate that the best method we identified can still accurately predict falls after accounting for missing data, we devised the

following experiment. For patients with sufficient data to make predictions the aforementioned model was used, and where individuals were missing the necessary data the majority prediction (no fall) was utilized instead, given that in reality, the vast majority of patients do not fall. The quantitative measures of predictive strength detailed above were also reported for this final model.

Results

339 patients were approached to participate in this study, of whom 54 fell during their in-patient stay. The demographics features for those two sub-groups (Fallers vs. Non-fallers) is described in Table 1. Figure 1 describes the reduction in sample size due to the relevant constraints (i.e. refusal to consent, inability to compete any tests, and specific contraindications for the Trail Making test). After accounting for all of the constraints, 211 individuals with demographic, patient reported outcome measure and Trail Making data remained. Five of the 211 individuals undertaking the Trail Making test had incomplete dataset due to administrator error in recording the resultant variables (Table 2).

The median time from admission to testing was 2 days (Range: 1 - 30). 71% (n = 229) of the population had all of the tests administered within 2 days of admission. There was no significant difference between the time from admission to testing when fallers and non-fallers were compared ($p = 0.27$). Age, number of years of formal education and ethnicity did not significantly differ between the faller and non-faller cohorts (Table 1). However, there were significantly more men ($p<0.05$) in the non-faller cohort and the vast majority of both groups identified as ethnically white (Table 1).

The primary raw scores for the Trail Making Test, time on part A and B, both demonstrated significant differences between the fallers and non-fallers, at the 0.01% significance levels (Table 2). The Trail Making error scores were significantly different for part B, but not part A (Table 2). The Trail Making composite score did not differ between the two groups (Table 2). Furthermore, of the three binary questions, only ‘having undergone surgery in the last month’ was not significantly associated with falling (Table 3). All 12 questions of the Walk-12 questionnaire differed significantly ($p < 0.01$) between fallers and non-fallers (Table 2).

The dataset was then restricted so that only individuals with data for all of the tests were included (i.e. restricting the dataset to those with complete Trails datasets, n = 206) to allow for fair and formal comparison between different models and sub-sets of the data. The modelling data (Table 3) suggests that the Trail Making test produces the best predictions (Wilcoxon signed-rank $p < .001$). Moreover, adding any of the other variables: demographic features, the binary questions, or physical function-related, did not significantly improve the models prediction capabilities (Wilcoxon signed-rank $p < .001$).

The three receiver operating characteristics curves (Fig. 2) demonstrate that the Trails data is the best predictor, and that the Random Forest is the best accompanying modelling method (Fig. 2). [Please insert figure 2]. The logistic regression method combined with the demographic data and the binary questions produces a reasonably good predictive model. However, when the Trails data is used instead of the demographic and binary data, but still using logistic regression method, the predictive power of the resulting model is significantly improved (Wilcoxon signed-rank on residuals $p < .001$). The result of the model can again be significantly improved (Wilcoxon signed-rank on residuals $p < .001$), by replacing the logistic regression method with the Random Forest (in combination with the trail making data). Thus, the combination of the Trail Making variables and the Random Forest appears to produce the best predictive model based on the available data. The best version of the model generated was capable of predicting with 68% (± 7.7) sensitivity, 90% (± 2.3) specificity, 0.600 (± 7.6) precision, and 0.630 (± 0.063) F1-score, in a population where the Trail Making data is available.

Finally, the unrestricted dataset with all individuals, including those without Trail Making data was utilized. After applying the majority classifier to all individuals without the Trail data, and employing the Random Forest and Trail Making for everyone else, the

predicted sensitivity was 51% and specificity was 94% (See Appendix Table S3).

Discussion

In this study we have used the Trail Making test in combination with the Random Forest to produce a falls prediction model. This model accurately identifies which tertiary neuroscience centre in-patients are at high risk of falling (sensitivity of 68% ± 7.7, and specificity of 90% ± 2.3). In our data set, neurological in-patients that fall are more likely to have impaired cognitive, and reduced self-reported physical function compared to those individuals that do not fall. To our knowledge this is the first study which demonstrates the applicability of machine learning methods when combined with cognitive data. However, external validation in a new sample is required before we can be certain of the veracity of these results, and therefore recommend it for use in clinical practice.

In a previous study, Kabeshova and colleagues demonstrated the superiority of machine learning predictive methods to classical predictive models in the falls prediction setting.[30] However, they used demographic and risk factor data, but not cognitive data. Our results suggest that the Trail Making test results are sufficient to predict falls risk in this sample, as the addition of demographic or physical function related variables did not improve predictive accuracy. Moreover, the Random Forest model appears to be these best statistical model to use in combination with the Trail Making test, similarly demonstrating the superiority of machine learning methods. Unfortunately, machine learning methods such as the Random Forest are characterized by their black-box nature, meaning that it is impossible to ascertain how and why they reach individual decisions. Consequently, we are unable to provide simple cut-off scores for different categories of risk as one might do if they were to create a simpler linear model.

As we alluded to in the introduction, previous studies have described the importance of several cognitive substrates to locomotion,[11,31] including processing speed (using the Digit

Symbol Substitution Test);[32] attention (based on dual task measurements);[33] and executive function.[12] Here we have utilized a single cognitive test, the Trail Making test, that some believe spans all of these areas of cognition (processing speed, attention, and executive function). The Trail Making test is widely accepted to be a test of executive function,[17] however its role as a test of attention and the speed of processing is less clear.

According to one theory of attention, by Zomeren & Spikman,[18] part A of the Trail Making test is a measure of the speed of processing, which they argue is a form of attentional processing, whereas part B is a measure of the tactical level of attention. As predicted by the theory, the primary variables (time taken to complete part A and time taken to complete B) differed significantly between fallers and non-fallers. Moreover, Zomeren & Spikman argue that errors on measures of the operational level of attention (i.e. Trail Making Part A) are not relevant to processing speed, and therefore should not differ between the two groups. Whereas, errors on tests of the tactical level of attention (i.e. Trail Making Part B) are relevant to the measurement of attention, and so should differ between the two groups.[18] The pattern of significant and non-significant error results in our data is consistent with these postulates.

The concordance of the results and theoretical postulates lends credibility to the suggestion that the Trail Making test may capture information from across both important cognitive substrates: attention (including speed of processing), and executive function. The implication for clinical practice is that collecting the Trail Making test alone may be sufficient for predicting falls, and the collecting of a comprehensive battery of tests is unnecessary. However, definitive evidence for this claim requires an additional study to demonstrate that adding the purpose built measures of attention and processing speed to the Trail Making data does not improve fall-related predictions.

Lack of methodological robustness has become one of the central criticism of medical prediction/prognostic research over the last decade.[34] A recent systematic review found that the falls prediction tool recommended by the UK's relevant statutory body is substantially less accurate than the original validation study suggested.[35,36] As such, the main strength of this study is the use of the gold standard statistical techniques, such as: cross validation to mimic new, unseen, data; and estimation of the errors associated with each prediction statistic,[37] to prevent overestimation of the generated models' predictive capabilities. For example, by removing the cross-validation stage in our model development method, instead of identifying one model, we find several combination of model type and data which each had sensitivity and specificity in excess of 90%, and utilize a variety of physical and demographic variables in addition to the Trail Making test. The result of omitting these methods is that spurious correlations specific only to the initially measured population are more likely to be retained by a model, and thus, the model becomes over-fitted, and over-fitted models are very unlikely to replicate their exceptional initial performance in subsequent replication/validation studies.

One of the main limitations of this study is a result of the data being collected in a single tertiary centre that covered acute neurological, neurosurgical, and neurorehabilitation care, suggesting that the generalization of these results should be considered carefully. Furthermore, the time to testing represents another potential limitation. In the 2 days (median) from admission to testing it is possible that the patients may have become cognitively fatigued, undergone procedures, or received medication that increased their risk of falling. Unfortunately, little can be done to mitigate this limitation in this study, as a reasonable period of time needs to be allowed for the patients to complete the outcome measures. Given that the validity of the model has been demonstrated in this

preliminary study, future attempts to use the model should be done at the time of admission, which would illustrate the effect, if any, the time to testing has on predicting the outcome. The other limitations of the study are those specific to our choice of outcome measures. The trail making test is known to suffer from practice effects, and the effects of previous testing with this particular measure were not corrected for in our analysis.[38] Moreover, the use of the Walk-12, instead of an objective measure of physical function, such as the 10m walk test, could also be seen as a limitation of the study, especially in light of recent evidence suggesting that the latter is more predictive of falls than the former.[39] As such, additional investigation into the combination of an objective physical function measure and a practice effect resistant cognitive test is required.

Clinical Messages

- Neurological inpatients that fall are more likely to have impaired cognitive and physical function compared to non-fallers.
- The Trail Making test is capable of accurately predicting falls in an in-patient neurological population.

Declarations –

Prof. Diane Playford was supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. The authors have no conflict of interests to declare. This study received no funding.

Contributions –

BAM and EDP conceived and planned the study with contribution from CD, DW and AS. BAM collected the data under the supervision of DW and EDP. Statistical analysis conducted by MB under the supervision of FJK, with contributions from BAM. Manuscript written by BAM and FJK, under the supervision of EDP.

Acknowledgements –

During the preparation of this manuscript for publication, unfortunately, Dr. Denise Waller died. We would like to acknowledge her significant personal and professional contributions to this project.

References

[1] – World Health Organization. WHO Global Report on Falls Prevention in Older Age. Ageing And Life Course, family And Community Health. WHO Press. 2007. Available at: http://www.who.int/ageing/publications/Falls_prevention7March.pdf [Accessed Jan. 2015].

[2] - Rubenstein LZ. Falls in older people: epidemiology, risk factors and strategies for prevention. *Age and ageing* 2006; 35: ii37-ii41.

[3] - Leal J, Gray AM, Prieto-Alhambra D, et al. Impact of hip fracture on hospital care costs: a population-based study. *Osteoporos Int.* 2016; 27: 549-58.

[4] - Stevens JA, Corso PS, Finkelstein EA, Miller TR. The costs of fatal and non-fatal falls among older adults. *Inj Prev.* 2006; 12: 290-5.

[5] - National Patient Safety Agency. Slips trips and falls in hospital, NPSA: London. 2007.

[6] - Healey F, Oliver D. Preventing falls and injury in hospitals. The evidence for interventions. *Health Care Risk Rep.* 2006; June: 12-7.

[7] - Nyberg L and Gustafson Y. Patient falls in stroke rehabilitation: a challenge to rehabilitation strategies. *Stroke* 1995; 26(5): 838–842.

[8] - Stevens JA, Corso PS, Finkelstein EA, Miller TR. The costs of fatal and non-fatal falls among older adults. Injury prevention. 2006 Oct 1;12(5):290-5.

[9] - Oliver D, Britton M, Seed P, Martin FC, Hopper AH. Development and evaluation of an evidence-based risk assessment tool (STRATIFY) to predict which elderly patients will fall: case control and cohort studies. *Br Med J*. 1997; 315: 1049-51.

[10] - Oliver D, Daly F, Martin FC, McMurdo ME. Risk factors and risk assessment tools for falls in hospital in-patients: a systematic review. Age and ageing. 2004 Mar 1;33(2):122-30.

[11] - Amboni M, Barone P, Hausdorff JM. Cognitive contributions to gait and falls: evidence and implications. *Mov Disord.* 2013; **28**: 1520-33.

[12] - Herman T, Mirelman A, Giladi N, Schweiger A, Hausdorff JM. Executive Control Deficits as a Prodrome to Falls in Healthy Older Adults: A Prospective Study Linking Thinking, Walking, and Falling. *J Gerontol A Bio. Sci Med Sci.* 2010; 65A: 1086-1092.

[13] - Mak M, Wong A, Pang M. Impaired Executive Function Can Predict Recurrent Falls in Parkinson's Disease. *Arch Phys Med Rehabil.* 2014: 95: 2390-2395.

[14] - Kalron A. The Relationship between Specific Cognitive Domains, Fear of Falling, and Falls in People with Multiple Sclerosis. *BioMed Research International.* 2014; Jul 24: 1-10.

[15] - Guaraldi P, Poda R, Calandra-Buonaura G, et al. Cognitive Function in Peripheral Autonomic Disorders. *PLoS ONE.* 2014; 9:e85020.

[16] - Lundin-Olsson, L., Nyberg, L. and Gustafson, Y. “Stops walking when talking” as a predictor of falls in elderly people. *Lancet.* 1997; 349: 617.

[17] - Strauss, E., Sherman, E., Spreen, O. and Spreen, O. (2006). *A compendium of neuropsychological tests*. Oxford: Oxford University Press.

[18] – Zomeren E, and Spikman J. Assessment of attention. In Halligan P, Kischka U, Marshall J, eds. *Oxford Handbook of Clinical Neuropsychology*. Oxford: Oxford University Press. 2003

[19] - Spikman JM, Kiers HA, Deelman BG, van Zomeren AH. Construct Validity of Concepts of Attention in Healthy Controls and Patients with CHI. *Brain and Cognition.* 2001; 47: 446-460.

[20] - Norman DA, Shallice T. Attention to action. In Consciousness and self-regulation 1986 (p. 1-18). Springer US.

[21] - Mazer B, Korner-Bitensky N, Sofer S. Predicting ability to drive after stroke. Arch Phys Med Rehabil. 1998;79(7):743–50. doi: 10.1016/S0003-9993(98)90350-1.

[22] - Classen S, Horgas A, Awadzi K, et al. Clinical predictors of older driver performance on a standardized road test. Traffic Inj Prev. 2008;9(5):456–62.

[23] - Army Individual Test Battery: Manual of directions and scoring. 1944. Washington, D. C. War Department, Adjutant General’s Office.

[24] - Reitan RM. Theoretical and methodological bases of the Halstead—Reitan Neuropsychological Test Battery. In I. Grant & K. M Adams (Eds.), *Neuropsychological assessment of neuropsychiatric disorders* (pp. 3–30). 1986. New York: Oxford University Press.

[25] - Holland A, O'Connor RJ, Thompson AJ, Playford ED, Hobart JC. Talking the talk on walking the walk. Journal of neurology. 2006 Dec 1;253(12):1594-602.

[26] - R Core Team. R: A language and environment for statistical computing. R. Foundation for Statistical Computing, Vienna, Austria. 2016. ISBN 3-900051-07-0, URL- http://www.R-project.org/.

[27] - Bischl B, Lang M, Kotthoff L, et al. mlr: Machine Learning in R. *Journal of Machine Learning Research.* 2016; **17**: 1-5.

[28] - Breiman L. Random forests. *Machine learning* 2001;45:5-32.

[29] - Office for National Statistics, (2011) Census: Aggregate data (England and Wales) [computer file]. UK Data Service Census Support. 2011. Downloaded from: http://infuse.mimas.ac.uk.

[30] - Kabeshova A, Launay CP, Gromov VA, et al. Falling in the elderly: Do statistical models matter for performance criteria of fall prediction? Results from two large population-based

studies. *European journal of internal medicine* 2016; **27**: 48-56.

[31] - Kearney FC, Harwood RH, Gladman JR, Lincoln N, Masud T. The Relationship between Executive Function and Falls and Gait Abnormalities in Older Adults: A Systematic Review. *Dementia and Geriatric Cognitive Disorders*. 2013; 36: 20-35.

[32] - Tamez E, Myerson J, Morris L, White DA, Baum C, Connor LT. Assessing executive abilities following acute stroke with the trail making test and digit span. *Behavioural neurology.* 2011; **24**: 177-85.

[33] - Ayers EI, Tow AC, Holtzer R, Verghese J. Walking while Talking and Falls in Aging. *Gerontology*. 2014; 60: 108-113.

[34] - Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, Altman DG, Moons KG. Reporting and methods in clinical prediction research: a systematic review. PLoS medicine. 2012 May 22;9(5):e1001221.

[35] - National Institute for Health and Care Excellence. *Falls. Assessment and prevention of falls in older people (full NICE guideline).* Clinical guideline 161. *National Institute for Health and Care Excellence*. 2013

[36] - Billington J, Fahey T, Galvin R. Diagnostic accuracy of the STRATIFY clinical prediction rule for falls: a systematic review and meta-analysis. *BMC family practice* 2012; **13**: 76.

[37] - Hastie, TJ, Tibshirani RJ, and Friedman JH. *The Elements Of Statistical Learning*. New York, NY: Springer, 2016. Print.

[38] - Buck KK, Atkinson TM, Ryan JP. Evidence of practice effects in variants of the Trail Making Test during serial assessment. Journal of clinical and experimental neuropsychology. 2008 Mar 7;30(3):312-8.

[39] - Allali G, Ayers EI, Verghese J. Multiple modes of assessment of gait are better than one to predict incident falls. Archives of gerontology and geriatrics. 2015 May 1;60(3):389-93.

Table 1 – Cohort Demographics

Demographic Data	*Fallers N = 54*	*Non-Fallers N = 283*
Sex		
Male	25 (46.3%)	169 (59.7 %)
Female	29 (53.7%)	114 (40.3 %)
Ethnicity*		
White	42 (77.8%)	219 (77.4%)
Asian/Asian British	5 (9.26%)	40 (14.2%)
Black/African/Caribbean/Black British	6 (11.1%)	13 (6.71%)
Other/Mixed	1 (1.85%)	5 (1.77%)
Age		
<19	0 (0.00%)	1 (0.35%)
19 - 29	6 (11.1%)	24 (8.48%)
29 - 39	3 (5.56%)	44 (15.5%)
39 - 49	10 (18.5%)	42 (14.8%)
49 - 59	14 (25.9%)	58 (20.5%)
59 - 69	10 (18.5%)	49 (17.3%)
69 - 79	6 (11.1%)	46 (16.3%)
79 - 89	5 (9.26%)	18 (6.36%)
89 – 99	0 (0.00%)	1 (0.35%)
Mean [95% Confidence Interval]	55.4 [50.7, 60.2]	54.7 [52.5, 56.9]
Primary Diagnosis		
Space Occupying Lesion	10	38
Under investigation / No known diagnosis	0	45
Spinal Cord Pathology	5	39
Stroke	15	24
Cephalgia (incl. migraine)	0	16
Intracranial Hypertension & Hydrocephalus	4	24
Disc-related Pathology	1	29
Extra-axial Haemorrhage	2	16
Cerebrovascular Malformation	3	9
Parkinsonism	2	9
Rapid Cognitive Decline	4	0
Myasthenia Gravis	1	3
CNS Vasculitis	2	2
Multiple Sclerosis	1	0
Inflammatory Encephalopathy	1	0
Gullian Barre Syndrome	1	0
Dropped Head Syndrome	1	0
Bilateral Progressive Optic Neuropathy	1	0

Conditions with 5 or less diagnosed individuals and no falls: Autonomic failure (5), Epilepsy (4), Cushing's Disease (3), Functional Motor Disorder (2), Tuberculosis (2), Motor Neuron Disease (1), Foot drop (1), Progressive Sensory Neuropathy (1), Polymyositis (1), Pneumocephalus (1), Phenylketonuria (1), Optic Neuritis (1), Neuromyelitis Optica (1), GAD-positive Ataxia Syndrome (1), Dystonic tremor (1), Dural Fistula (1), Chronic fatigue syndrome (1), Back and leg pain (1).		
Years of Education^		
Mean [95% Confidence Interval]	13.1 [12.1, 14.1]	13.4 [12.9, 14.0]

*Ethnicity reported in line with the standardized classification used by the office for national statistics.[29] ^ Total number of years in primary, secondary, further &/or higher education. # - Significant difference (P <0.01)

Table 2 -Non-Faller and Faller Summary Statistics for the Trail Making Test Variables

Test	Population	Sample Size (Participants)	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Fallers vs. Non-fallers Significance (t-test p value)
Time to Complete Part A (Seconds)	Faller	39	15.0	49.0	76.0	80.43	90.5	300.0	3.0×10^{-5}
	Non-Faller	172	14.0	26.0	34.0	42.51	48.0	131.0	
Number of Errors - Part A	Faller	38	0.0	0.0	0.0	1.10	1.0	3.0	1.2×10^{-2}
	Non-Faller	172	0.0	0.0	0.0	0.93	0.0	2.0	
Time to complete Part B (Seconds)	Faller	39	42.0	176.0	253.0	200.71	294.5	300.0	3.9×10^{-8}
	Non-Faller	171	32.0	84.0	131.0	121.57	191.0	300.0	
Number of Errors - Part B	Faller	38	0.0	0.3	2.0	1.27	3.0	8.0	2.0×10^{-4}
	Non-Faller	168	0.0	0.0	0.0	0.80	1.0	7.0	
Time to Complete Part B / Time to Complete Part A	Faller	38	1.0	2.4	2.8	2.79	4.2	10.5	4.3×10^{-1}
	Non-Faller	171	1.6	2.5	3.6	2.97	4.6	7.9	

Table 3 – Non-Faller and Faller Summary for the Walk-12 Questions

Test	Population	Sample Size (Participants)	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Fallers vs Non-Fallers Significance (t-test p value)
Question 1	Faller	50	0.0	2.3	4.0	3.88	5.0	5.0	9.5×10^{-3}
	Non-Faller	204	0.0	1.0	3.0	2.91	4.0	5.0	
Question 2	Faller	50	0.0	2.0	5.0	4.28	5.0	5.0	1.8×10^{-1}
	Non-Faller	204	0.0	1.0	4.0	3.34	5.0	5.0	
Question 3	Faller	50	0.0	2.0	4.0	3.81	5.0	5.0	2.3×10^{-2}
	Non-Faller	204	0.0	1.0	3.0	2.89	4.0	5.0	
Question 4	Faller	50	0.0	2.0	4.0	3.53	5.0	5.0	1.2×10^{-1}
	Non-Faller	204	0.0	1.0	3.0	2.95	4.0	5.0	
Question 5	Faller	50	0.0	2.0	3.5	3.53	5.0	5.0	2.6×10^{-1}
	Non-Faller	204	0.0	1.8	3.0	2.87	4.0	5.0	
Question 6	Faller	50	0.0	3.0	4.5	4.19	5.0	5.0	6.7×10^{-2}
	Non-Faller	204	0.0	2.0	3.0	3.29	5.0	5.0	
Question 7	Faller	50	0.0	2.3	4.5	4.13	5.0	5.0	4.9×10^{-2}
	Non-Faller	204	0.0	2.0	3.0	3.17	4.0	5.0	
Question 8	Faller	50	0.0	1.0	4.0	3.72	5.0	5.0	1.8×10^{-1}
	Non-Faller	204	0.0	1.0	3.0	2.83	5.0	5.0	
Question 9	Faller	50	0.0	1.0	4.0	3.91	5.0	5.0	6.1×10^{-2}
	Non-Faller	204	0.0	1.0	2.0	2.68	5.0	5.0	
Question 10	Faller	50	0.0	3.0	5.0	4.13	5.0	5.0	6.5×10^{-2}
	Non-Faller	204	0.0	2.0	3.0	3.27	5.0	5.0	
Question 11	Faller	50	0.0	3.0	4.0	4.09	5.0	5.0	6.7×10^{-2}
	Non-Faller	204	0.0	1.0	3.0	3.14	5.0	5.0	
Question 12	Faller	50	0.0	4.0	5.0	4.44	5.0	5.0	8.7×10^{-3}
	Non-Faller	204	0.0	1.0	4.0	3.25	5.0	5.0	

Table 4 – Non-Faller and Faller Summary Statistics for the Three Binary Questions

Test	Population	Sample Size (Participants)	Yes	No	Fallers vs. Non-Fallers Significance (Chi-squared test p value)
Undergone neurosurgery in the last month?	Faller	54	31	23	4.6×10^{-1}
	Non-Faller	283	144	139	
Fallen in the last month?	Faller	54	29	25	4.6×10^{-4}
	Non-Faller	283	80	203	
Experienced a change in physical function change in the last month?	Faller	54	43	11	1.8×10^{-3}
	Non-Faller	283	158	125	

Table 5: Best possible prediction from the five different variable sets.

Data Utilized	Best Method	Mean Misclassification Error (MMCE)	Sensitivity	Specificity	Precision	F1 - Score
Trail	Random Forest	0.117 (± .022)	0.550 (± .083)	0.958 (± .015)	0.758 (± .085)	0.619 (± .071)
Walk-12	Linear Discriminant Analysis	0.169 (±0.024)	0.100 (±0.045)	0.990 (±0.007)	0.700 (±0.230)	0.153 (±0.074)
Demographics	SVM (Gauss)	0.139 (±0.019)	0.153 (±0.049)	0.996 (±0.004)	0.833 (±0.118)	0.231 (±0.075)
Trail + Walk-12	Random Forest	0.143 (±0.027)	0.450 (±0.089)	0.955 (±0.018)	0.722 (±0.103)	0.516 (±0.084)
Trail + Demographics	avNNet	0.132 (±0.025)	0.575 (±0.081)	0.928 (±0.020)	0.671 (±0.082)	0.553 (±0.067)

The data set upon which the following table is based was the restricted data set consisting of those with trail data (excluding those for which the trail data was missing), i.e. n = 206, of the total 337.

Figure 1(A) -

339 patients were approached between November 17th and December 17th 2014

Exclude all patients who declined to consent → *2 patients excluded*

337 patients remaining

Exclude all patients who not complete the test battery, but consented for their medical notes to be read to answer the three binary questions and identify the relevant demographic data → 83 patients excluded: 29 patients were not fluent in English and therefore testing was deemed inappropriate given the lack of language proficient administrators, and valid translations of the tools. The medical teams of 37 patients deemed testing inappropriate due to severe cognitive impairment, communication difficulties, or severe mood/behavioral problems. Seventeen patients agreed to their data

254 patients remaining

Exclude all patients who were unable to complete the Patient Reported Outcome Measure → *0 patients excluded: all 254 patients completed the patient reported outcome measure, as it was the least onerous of the tasks*

254 patients remaining

Exclude all patients that fall into the exclusion criteria for the Trails test (specific contra-indication for each test that have been highlighted in appendix Table 1) → *43 patients excluded*

211 patients remaining

Exclude all patients with any missing data → *5 patients excluded (Missing Trail Making Data due to administrator error.*

206 patients remaining

Figure 1 – A Flowchart to illustrate Sample Size Constraints

The diagram describes the number of individuals that were excluded from the analysis due to the different constraints surrounding consent, data collection and appropriateness of the tests in the context of the individual's diagnosis and/or deficits. In total 339 patients were approached, which resulted in a final sample of 211 cases after accounting for the aforementioned constraints. The initial modelling was conducted on this sample of 206 individuals. An additional set of experiments was conducted to demonstrate how the final model would perform in the real world (on all 337 consenting participants), as exclusion is not a viable solution in the clinical context, but was deemed acceptable for creating the model.

Figure 2

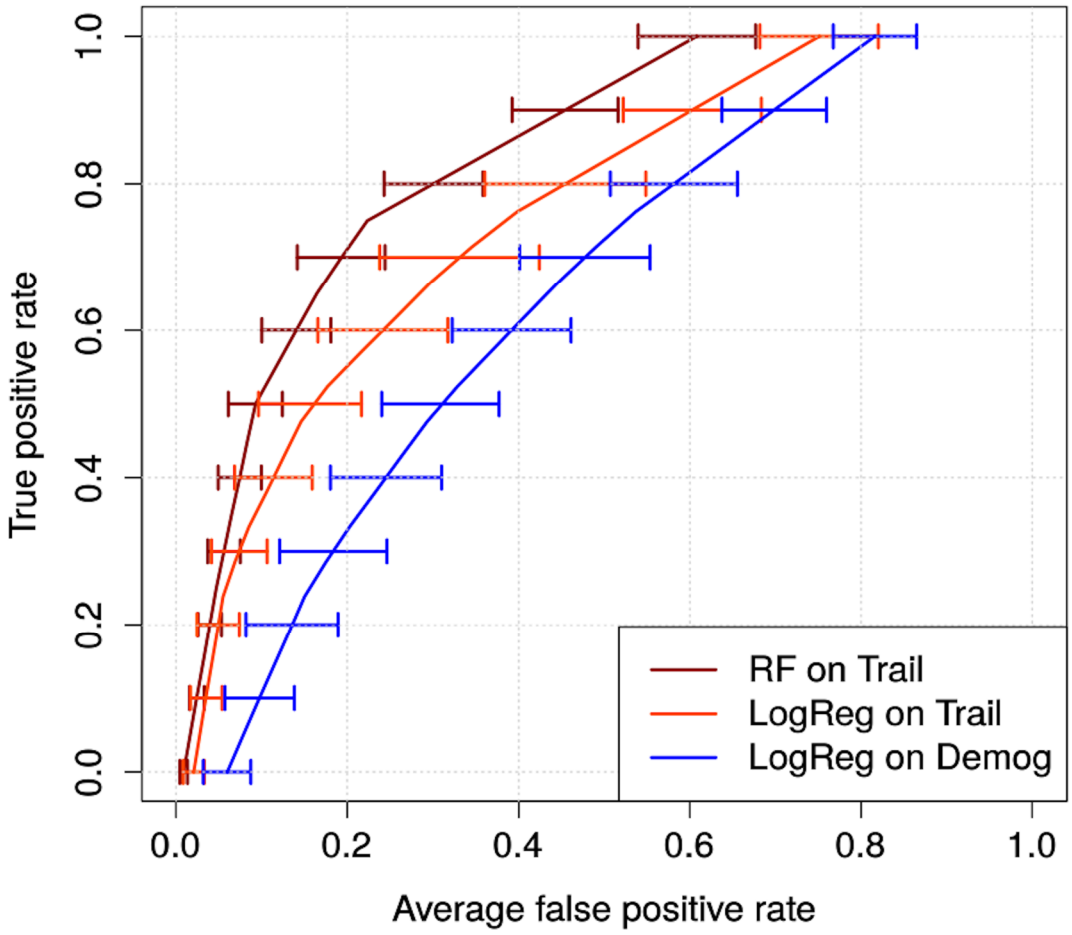
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Average false positive rate
0.0
0.2
0.4
0.6
0.8
1.0
RF on Trail
LogReg on Trail
LogReg on Demog

Figure 2 – The Receiver Operating Characteristics (ROC) for Random Forest and Logistic Regression based classifiers

The data upon which the following ROCs are based is the restricted data set consisting of those individuals with trail data (i.e. excluding those for which the trail data was missing). The figure illustrates the conclusion that the random forest (RF) based predictor appears to be superior to that of logistic regression (LogReg) when both utilize only the Trails data ($p < 0.001$). Moreover, both of these models are superior to the baseline model of demographic (Demog +Binary) data ($p < 0.001$), consisting of common risk factors for falls) and the logistic regression model, which suggests that using a test of cognitive function (i.e. the Trail Making Test) appears to improve predictive capabilities, at least in our dataset. The Area under the ROCs (AUROCs) are LogReg on Demog + Binary (0.65), LogReg on Trail (0.78), and RF on Trail (0.87).